

Stateology: State-Level Interactive Charting of Language, Feelings, and Values

Konstantinos Pappas, Steven Wilson, and Rada Mihalcea

Computer Science and Engineering

University of Michigan

Ann Arbor, MI 48109

{pappus, steverw, mihalcea}@umich.edu

Abstract

People’s personality and motivations are manifest in their everyday language usage. With the emergence of social media, ample examples of such usage are procurable. In this paper, we aim to analyze the vocabulary used by close to 200,000 Blogger users in the U.S. with the purpose of geographically portraying various demographic, linguistic, and psychological dimensions at the state level. We give a description of a web-based tool for viewing maps that depict various characteristics of the social media users as derived from this large blog dataset of over two billion words.

1 Introduction

Blogging gained momentum in 1999 and became especially popular after the launch of freely available, hosted platforms such as `blogger.com` or `livejournal.com`. Blogging has progressively been used by individuals to share news, ideas, and information, but it has also developed a mainstream role to the extent that it is being used by political consultants and news services as a tool for outreach and opinion forming as well as by businesses as a marketing tool to promote products and services (Nardi et al., 2004).

For this paper, we compiled a very large geolocated collection of blogs, written by individuals located in the U.S., with the purpose of creating insightful mappings of the blogging community. In particular, during May-July 2015, we gathered the profile information for all the users that have self-reported their location in the U.S., along with a number of posts for all their associated blogs. We uti-

lize this blog collection to generate maps of the U.S. that reflect user demographics, language use, and distributions of psycholinguistic and semantic word classes. We believe that these maps can provide valuable insights and partial verification of previous claims in support of research in linguistic geography (Brice, 2003), regional personality (Rogers and Wood, 2010), and language analysis (Schwartz et al., 2013; Schmitt et al., 2007), as well as psychology and its relation to human geography (Kitchin et al., 1997).

2 Data Collection

Our premise is that we can generate informative maps using geolocated information available on social media; therefore, we guide the blog collection process with the constraint that we only accept blogs that have specific location information. Moreover, we aim to find blogs belonging to writers from all 50 U.S. states, which will allow us to build U.S. maps for various dimensions of interest.

We first started by collecting a set of profiles of bloggers that met our location specifications by searching individual states on the profile finder on `http://www.blogger.com`. Starting with this list, we can locate the profile page for a user, and subsequently extract additional information, which includes fields such as name, email, occupation, industry, and so forth. It is important to note that the profile finder only identifies users that have an exact match to the location specified in the query; we thus built and ran queries that used both state abbreviations (e.g., TX, AL), as well as the states’ full names (e.g., Texas, Alabama).

After completing all the processing steps, we

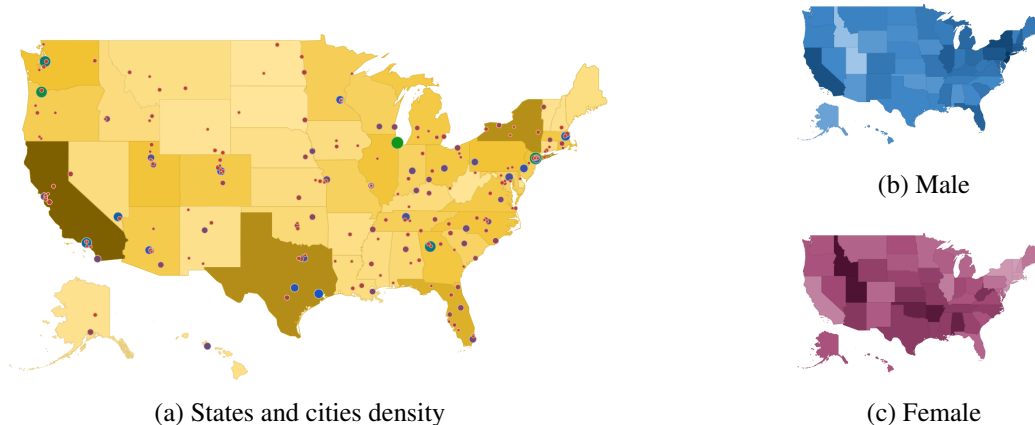


Figure 1: Geographical distribution of bloggers in the 50 U.S. states.

identified 197,527 bloggers with state location information. For each of these bloggers, we found their blogs (note that a blogger can have multiple blogs), for a total of 335,698 blogs. For each of these blogs, we downloaded the 21 most recent blog postings, which were cleaned of HTML tags and tokenized, resulting in a collection of 4,600,465 blog posts.

3 Maps from Blogs

Our dataset provides mappings between location, profile information, and language use, which we can leverage to generate maps that reflect demographic, linguistic, and psycholinguistic properties of the population represented in the dataset.¹

3.1 People Maps

The first map we generate depicts the distribution of the bloggers in our dataset across the U.S. Figure 1a shows the density of users in our dataset in each of the 50 states. For instance, the densest state was found to be California with 11,701 users. The second densest is Texas, with 9,252 users, followed by New York, with 9,136. The state with the fewest bloggers is Delaware with 1,217 users. Not surprisingly, this distribution correlates well with the population of these states,² with a Spearman’s rank correlation ρ of 0.91 and a p-value < 0.0001 , and is very similar to the one reported in Lin and Halavais (2004).

¹In all the maps we generate, the darker the color of a state, the higher the proportion of instances in that state that match the criterion used to generate the map.

²<http://www.census.gov/2010census/data/apportionment-dens-text.php>

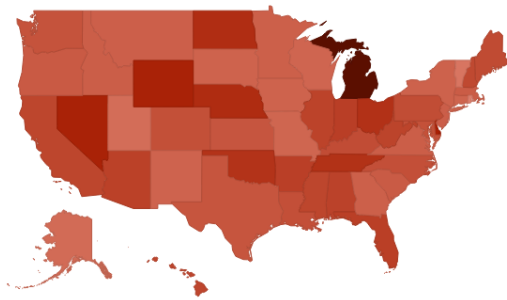
Figure 1a also shows the cities mentioned most often in our dataset. In particular, it illustrates all 227 cities that have at least 100 bloggers. The bigger the dot on the map, the larger the number of users found in that city. The five top blogger-dense cities, in order, are: Chicago, New York, Portland, Seattle, and Atlanta.

We also generate two maps that delineate the gender distribution in the dataset. Overall, the blogging world seems to be dominated by females: out of 153,209 users who self-reported their gender, only 52,725 are men and 100,484 are women. Figures 1b and 1c show the percentage of male and female bloggers in each of the 50 states. As seen in this figure, there are more than the average number of male bloggers in states such as California and New York, whereas Utah and Idaho have a higher percentage of women bloggers.

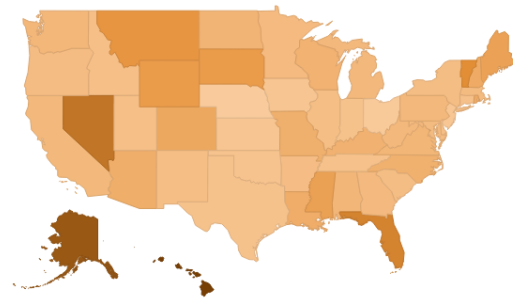
Another profile element that can lead to interesting maps is the *Industry* field (Holmes and Stevens, 2004). Using this field, we created different maps that plot the geographical distribution of industries across the country. As an example, Figure 2 shows the percentage of the users in each state working in the automotive and tourism industries respectively.

3.2 Linguistic Maps

Another use of the information found in our dataset is to build linguistic maps, which reflect the geographic lexical variation across the 50 states (Eisenstein et al., 2010). We generate maps that represent the relative frequency by which a word occurs in the different states. Figure 3 shows sample maps

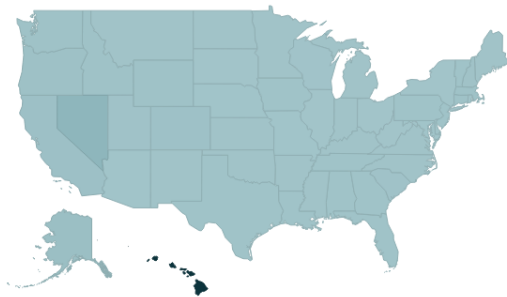


Automotive

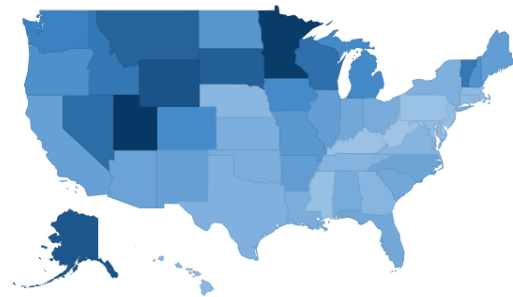


Tourism

Figure 2: Industry distribution across the 50 U.S. states for two selected industries.



Maui



Lake

Figure 3: Word distributions across the 50 U.S. states for two selected words.

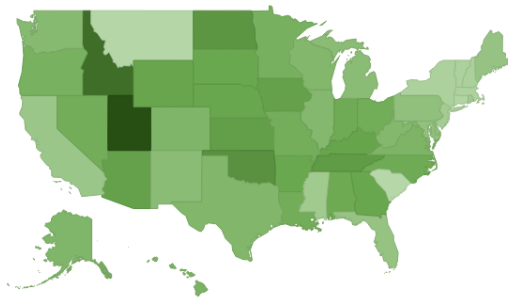
created for two different words. The figure shows the map generated for one location specific word, *Maui*, which unsurprisingly is found predominantly in Hawaii, and a map for a more common word, *lake*, which has a high occurrence rate in Minnesota (Land of 10,000 Lakes) and Utah (home of the Great Salt Lake). Our demo described in Section 4, can also be used to generate maps for function words, which can be very telling regarding people’s personality (Chung and Pennebaker, 2007).

3.3 Psycholinguistic and Semantic Maps

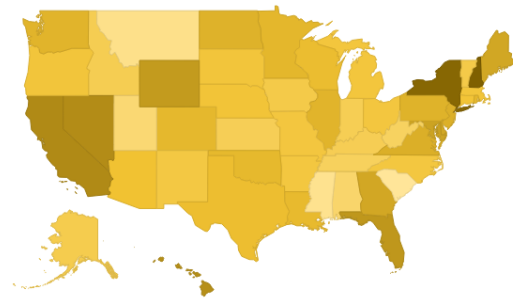
LIWC. In addition to individual words, we can also create maps for word categories that reflect a certain psycholinguistic or semantic property. Several lexical resources, such as Roget or Linguistic Inquiry and Word Count (Pennebaker et al., 2001), group words into categories. Examples of such categories are **MONEY**, which includes words such as remuneration, dollar, and payment; or **POSITIVE FEELINGS** with words such as happy, cheerful, and celebration. Using the distribution of the individual words in a

category, we can compile distributions for the entire category, and therefore generate maps for these word categories. For instance, figure 4 shows the maps created for two categories: **POSITIVE FEELINGS** and **MONEY**. The maps are not surprising, and interestingly they also reflect an inverse correlation between **MONEY** and **POSITIVE FEELINGS**.

Values. We also measure the usage of words related to people’s core values as reported by Boyd et al. (2015). The sets of words, or themes, were excavated using the Meaning Extraction Method (MEM) (Chung and Pennebaker, 2008). MEM is a topic modeling approach applied to a corpus of texts created by hundreds of survey respondents from the U.S. who were asked to freely write about their personal values. To illustrate, Figure 5 shows the geographical distributions of two of these value themes: **RELIGION** and **HARD WORK**. Southeastern states often considered as the nation’s “Bible Belt” (Heatwole, 1978) were found to have generally higher usage of **RELIGION** words such as *God*, *bible*, and *church*. Another broad trend was that western-

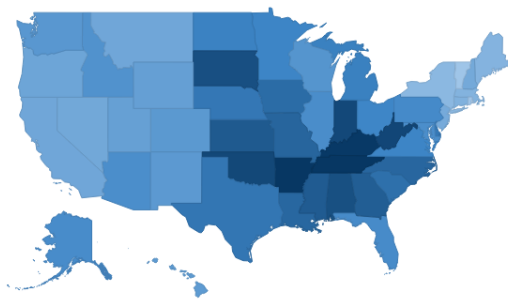


Positive feelings

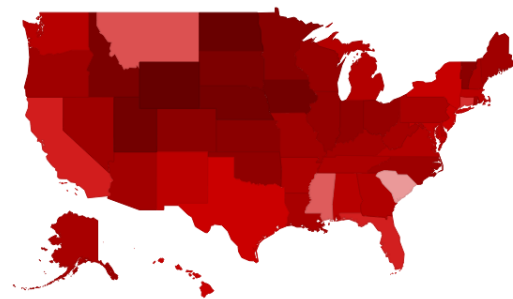


Money

Figure 4: LIWC distributions across the 50 U.S. states for two selected semantic categories.



Religion



Hard work

Figure 5: Values distributions across the 50 U.S. states for two selected values.

central states (e.g., Wyoming, Nebraska, Iowa) commonly blogged about **HARD WORK**, using words such as *hard*, *work*, and *job* more often than bloggers in other regions.

4 Web Demonstration

A prototype, interactive charting demo is available at <http://lit.eecs.umich.edu/~geoliwc/>. In addition to drawing maps of the geographical distributions on the different LIWC categories, the tool can report the three most and least correlated LIWC categories in the U.S.³ and compare the distributions of any two categories.

5 Conclusions

In this paper, we showed how we can effectively leverage a prodigious blog dataset. Not only does the dataset bring out the extensive linguistic content reflected in the blog posts, but also includes location information and rich metadata. These data al-

low for the generation of maps that reflect the demographics of the population, variations in language use, and differences in psycholinguistic and semantic categories. These mappings can be valuable to both psychologists and linguists, as well as lexicographers. A prototype demo has been made available together with the code used to collect our dataset.⁴

Acknowledgments

This material is based in part upon work supported by the National Science Foundation (#1344257) and by the John Templeton Foundation (#48503). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the John Templeton Foundation. We would like to thank our colleagues Hengjing Wang, Jiatao Fan, Xinghai Zhang, and Po-Jung Huang who provided technical help with the implementation of the demo.

³We use the Spearman rank correlation coefficient to calculate correlation.

⁴<http://lit.eecs.umich.edu/downloads.html>

References

- [Boyd et al.2015] Ryan L Boyd, Steven R Wilson, James W Pennebaker, Michal Kosinski, David J Stillwell, and Rada Mihalcea. 2015. Values in words: Using language to evaluate and understand personal values. In *Ninth International AAAI Conference on Web and Social Media*.
- [Brice2003] William C Brice. 2003. The geography of language. *Companion Encyclopedia of Geography: The Environment and Humankind*, page 107.
- [Chung and Pennebaker2007] Cindy Chung and James W Pennebaker. 2007. The psychological functions of function words. *Social communication*, pages 343–359.
- [Chung and Pennebaker2008] Cindy K Chung and James W Pennebaker. 2008. Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of Research in Personality*, 42(1):96–132.
- [Eisenstein et al.2010] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics.
- [Heatwole1978] Charles A Heatwole. 1978. The bible belt: A problem in regional definition. *Journal of Geography*, 77(2):50–55.
- [Holmes and Stevens2004] Thomas J Holmes and John J Stevens. 2004. Spatial distribution of economic activities in north america. *Handbook of regional and urban economics*, 4:2797–2843.
- [Kitchin et al.1997] Robert M Kitchin, Mark Blades, and Reginald G Golledge. 1997. Relations between psychology and geography. *Environment and Behavior*, 29(4):554–573.
- [Lin and Halavais2004] Jia Lin and Alexander Halavais. 2004. Mapping the blogosphere in america. In *Workshop on the weblogging ecosystem, 13th international world wide web conference*.
- [Nardi et al.2004] Bonnie A. Nardi, Diane J. Schiano, and Michelle Gumbrecht. 2004. Blogging as social activity, or, would you let 900 million people read your diary? In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work, CSCW '04*, pages 222–231, New York, NY, USA. ACM.
- [Pennebaker et al.2001] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001.
- [Rogers and Wood2010] Katherine H Rogers and Dustin Wood. 2010. Accuracy of united states regional personality stereotypes. *Journal of Research in Personality*, 44(6):704–713.
- [Schmitt et al.2007] David P Schmitt, Jüri Allik, Robert R McCrae, and Verónica Benet-Martínez. 2007. The geographic distribution of big five personality traits patterns and profiles of human self-description across 56 nations. *Journal of cross-cultural psychology*, 38(2):173–212.
- [Schwartz et al.2013] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.